Теория систем массового обслуживания

Теория массового обслуживания опирается на теорию вероятностей и математическую статистику.

При исследовании операций часто приходится сталкиваться с системами, предназначенными для многоразового использования при решении однотипных задач. Возникающие при этом процессы получили название ***процессов обслуживания***, а системы — ***систем массового обслуживания (СМО)***. Примерами таких систем являются телефонные системы, ремонтные мастерские, вычислительные комплексы, билетные кассы, магазины, парикмахерские и т.п.

Каждая СМО состоит из определенного числа обслуживающих единиц (приборов, устройств, пунктов, станций), которые будем называть ***каналами обслуживания***. Каналами могут быть линии связи, рабочие точки, вычислительные машины, продавцы и др. По числу каналов СМО подразделяют на ***одноканальные*** и ***многоканальные***.

Заявки поступают в СМО обычно не регулярно, а случайно, образуя так называемый ***случайный поток заявок (требований)***. Обслуживание заявок, вообще говоря, также продолжается какое-то случайное время. Случайный характер потока заявок и времени обслуживания приводит к тому, что СМО оказывается загруженной неравномерно: в какие-то периоды времени скапливается очень большое количество заявок (они либо становятся в очередь, либо покидают СМО необслуженными), в другие же периоды СМО работает с недогрузкой или простаивает.

Предметом теории массового обслуживания является построение математических моделей, связывающих заданные условия работы СМО (число каналов, их производительность, характер потока заявок и т.п.) с показателями эффективности СМО, описывающими ее способность справляться с потоком заявок.

В качестве ***показателей эффективности СМО*** используются: среднее число заявок, обслуживаемых в единицу времени; среднее число заявок в очереди; среднее время ожидания обслуживания; вероятность отказа в

обслуживании без ожидания; вероятность того, что число заявок в очереди превысит определенное значение и т.п.

СМО делят на два основных типа (класса): ***СМО с отказами*** и ***СМО с ожиданием (очередью)***. В СМО с отказами заявка, поступившая в момент, когда все каналы заняты, получает отказ, покидает СМО и в дальнейшем процессе обслуживания не участвует (например, заявка на телефонный разговор в момент, когда все каналы заняты, получает отказ и покидает СМО необслуженной). В СМО с ожиданием заявка, пришедшая в момент, когда все каналы заняты, не уходит, а становится в очередь на обслуживание.

СМО с ожиданием подразделяются на разные виды в зависимости от того, как организована очередь: с ограниченной или неограниченной длиной очереди, с ограниченным временем ожидания и т.п.

Для классификации СМО важное значение имеет ***дисциплина обслуживания***, определяющая порядок выбора заявок из числа поступивших и порядок распределения их между свободными каналами. По этому признаку обслуживание заявки может быть организовано по принципу "первая пришла — первая обслужена", "последняя пришла — первая обслужена" (такой порядок может применяться, например, при извлечении для обслуживания изделий со склада, ибо последние из них оказываются часто более доступными) или обслуживание с приоритетом (когда в первую очередь обслуживаются наиболее важные заявки). Приоритет может быть как *абсолютным*, когда более важная заявка"вытесняет" из-под обслуживания обычную заявку (например, в случае аварийной ситуации плановые работы ремонтных бригад прерываются до ликвидации аварии), так и *относительным*, когда более важная заявка получает лишь "лучшее" место в очереди.

Процесс работы СМО представляет собой ***случайный процесс***.

Под ***случайным (вероятностным или стохастическим) процессом*** понимается процесс изменения во времени состояния какой-либо системы в соответствии с вероятностными закономерностями.

Процесс называется ***процессом с дискретными состояниями***, если его возможные состояния $S1,S2,\ldots,Sn$ можно заранее перечислить, а переход

системы из состояния в состояние происходит мгновенно (скачком). Процесс называется ***процессом с непрерывным временем***, если моменты возможных переходов системы из состояния в состояние не фиксированы заранее, а случайны.

Процесс работы СМО представляет собой случайный процесс с дискретными состояниями и непрерывным временем. Это означает, что состояние СМО меняется скачком в случайные моменты появления каких-то событий (например, прихода новой заявки, окончания обслуживания и т.п.).

Математический анализ работы СМО существенно упрощается, если процесс этой работы — марковский. Случайный процесс называется ***марковским*** или ***случайным процессом без последствия***, если для любого момента времени t0 вероятностные характеристики процесса в будущем зависят только от его состояния в данный момент t0 и не зависят от того, когда и как система пришла в это состояние.

Пример марковского процесса: система S — счетчик в такси. Состояние системы в момент t характеризуется числом километров (десятых долей километров), пройденных автомобилем до данного момента. Пусть в момент t0 счетчик показывает S0. Вероятность того, что в момент t>t0 счетчик покажет то или иное число километров (точнее, соответствующее число рублей) S1, зависит от S0, но не зависит от того, в какие моменты времени изменялись показания счетчика до момента t0.

Многие процессы можно приближенно считать марковскими. Например, процесс игры в шахматы; система S — группа шахматных фигур. Состояние системы характеризуется числом фигур противника, сохранившихся на доске в момент t0. Вероятность того, что в момент t>t0 материальный перевес будет на стороне одного из противников, зависит в первую очередь от того, в каком состоянии находится система в данный момент t0, а не от того, когда и в какой последовательности исчезли фигуры с доски до момента t0.

В ряде случаев предысторией рассматриваемых процессов можно просто пренебречь и применять для их изучения марковские модели.

При анализе случайных процессов с дискретными состояниями удобно пользоваться геометрической схемой — так называемым ***графом состояний***. Обычно состояния системы изображаются прямоугольниками (кружками), а возможные переходы из состояния в состояние — стрелками (ориентированными дугами), соединяющими состояния.

Для математического описания марковского случайного процесса с дискретными состояниями и непрерывным временем, протекающего в СМО, познакомимся с одним из важных понятий теории вероятностей — понятием потока событий.

Под ***потоком событий*** понимается последовательность однородных событий, следующих одно за другим в какие-то случайные моменты времени (например, поток вызовов на телефонной станции, поток отказов ЭВМ, поток покупателей и т.п.).

Поток характеризуется ***интенсивностью*** λ — частотой появления событий или средним числом событий, поступающих в СМО в единицу времени.

Поток событий называется ***регулярным***, если события следуют одно за другим через определенные равные промежутки времени. Например, поток изделий на конвейере сборочного цеха (с постоянной скоростью движения) является регулярным.

Поток событий называется ***стационарным***, если его вероятностные характеристики не зависят от времени. В частности, интенсивность стационарного потока есть величина постоянная: $\lambda(t)=\lambda$. Например, поток автомобилей на городском проспекте не является стационарным в течение суток, но этот поток можно считать стационарным в течение суток, скажем, в часы пик. Обращаем внимание на то, что в последнем случае фактическое число проходящих автомобилей в единицу времени (например, в каждую минуту) может заметно отличаться друг от друга, но среднее их число будет постоянно и не будет зависеть от времени.

Поток событий называется ***потоком без последействия***, если для любых двух непересекающихся участков времени $\tau 1$ и $\tau 2$ — число событий,

попадающих на один из них, не зависит от числа событий, попадающих на другие. Например, поток пассажиров, входящих в метро, практически не имеет последействия. А, скажем, поток покупателей, отходящих с покупками от прилавка, уже имеет последействие (хотя бы потому, что интервал времени между отдельными покупателями не может быть меньше, чем минимальное время обслуживания каждого из них).

Поток событий называется ***ординарным***, если вероятность попадания на малый (элементарный) участок времени Δt двух и более событий пренебрежимо мала по сравнению с вероятностью попадания одного события. Другими словами, поток событий ординарен, если события появляются в нем поодиночке, а не группами. Например, поток поездов, подходящих к станции, ординарен, а поток вагонов не ординарен.

Поток событий называется ***простейшим*** *(или* ***стационарным пуассоновским****), если он одновременно стационарен, ординарен и не имеет последействия.* Название "простейший" объясняется тем, что СМО с простейшими потоками имеет наиболее простое математическое описание. Заметим, что регулярный поток не является "простейшим", так как он обладает последействием: моменты появления событий в таком потоке жестко зафиксированы.

Простейший поток в качестве предельного возникает в теории случайных процессов столь же естественно, как в теории вероятностей нормальное распределение получается в качестве предельного для суммы случайных величин: *при наложении (суперпозиции) достаточно большого числа* n *независимых, стационарных и ординарных потоков (сравнимых между собой по интенсивностям* λi $(i=1,2,\ldots,n)$ *получается поток, близкий к простейшему с интенсивностью* λ*, равной сумме интенсивностей входящих потоков, т.е.* $\lambda=\sum_{i=1}^{n}\lambda i$. *Рассмотрим на оси времени* Ot *простейший поток событий как неограниченную последовательность случайных точек.*

Можно показать, что для простейшего потока число т событий (точек), попадающих на произвольный участок времени τ, распределено по ***закону Пуассона*** $Pm(\tau)=((\lambda\tau)^m/m!)e^{-\lambda\tau}$, для которого математическое ожидание случайной величины равно ее дисперсии: $a=\sigma 2=\lambda\tau$.

В частности, вероятность того, что за время τ не произойдет ни одного события ($m=0$), равна $P0(\tau)=e-\lambda\tau$. Найдем распределение интервала времени T между произвольными двумя соседними событиями простейшего потока. Вероятность того, что на участке времени длиной t не появится ни одного из последующих событий, равна $P(T\geqslant t)=e-\lambda t$, а вероятность противоположного события, т.е. функция распределения случайной величины T, есть $F(t)=P(T<t)=1-e-\lambda t$.

Плотность вероятности случайной величины есть производная ее функции распределения, т.е. $\varphi(t)=F'(t)=\lambda e-\lambda t$.

Распределение, задаваемое плотностью вероятности или функцией распределения, называется ***показательным*** (или ***экспоненциальным***). Таким образом, интервал времени между двумя соседними произвольными событиями имеет показательное распределение, для которого математическое ожидание равно среднему квадратическому отклонению случайной величины $a=\sigma=1\lambda$ и обратно по величине интенсивности потока λ.

Важнейшее свойство показательного распределения (присущее только показательному распределению) состоит в следующем: если промежуток времени, распределенный по показательному закону, уже длился некоторое время τ, то это никак не влияет на закон распределения оставшейся части промежутка $(T-\tau)$: он будет таким же, как и закон распределения всего промежутка T.

Другими словами, для интервала времени T между двумя последовательными соседними событиями потока, имеющего показательное распределение, любые сведения о том, сколько времени протекал этот интервал, не влияют на закон распределения оставшейся части. Это свойство показательного закона представляет собой, в сущности, другую формулировку для "отсутствия последействия" — основного свойства простейшего потока.

Для простейшего потока с интенсивностью λ вероятность попадания на ***элементарный (малый)*** отрезок времени Δt хотя бы одного события потока равна согласно $P\Delta t = P(T < \Delta t) = 1 - e^{-\lambda \Delta t} \approx \lambda \Delta t$.

Заметим, что эта приближенная формула, получаемая заменой функции $e^{-\lambda \Delta t}$ лишь двумя первыми членами ее разложения в ряд по степеням Δt, тем точнее, чем меньше Δt).

На первичное развитие теории массового обслуживания оказали особое влияние работы датского ученого А.К. Эрланга (1878-1929).

Теория массового обслуживания – область прикладной математики, занимающаяся анализом процессов в системах производства, обслуживания, управления, в которых однородные события повторяются многократно, например, на предприятиях бытового обслуживания; в системах приема, переработки и передачи информации; автоматических линиях производства и др.

Предметом теории массового обслуживания является установление зависимостей между характером потока заявок, числом каналов обслуживанбия, производительностью отдельного канала и эффективным обслуживанием с целью нахождения наилучших путей управления этими процессами.

Задача теории массового обслуживания – установить зависимость результирующих показателей работы системы массового обслуживания (вероятности того, что заявка будет обслужена; математического ожидания числа обслуженных заявок и т.д.) от входных показателей (количества каналов в системе, параметров входящего потока заявок и т.д.). Результирующими показателями или интересующими нас характеристиками СМО являются – показатели эффективности СМО, которые описывают способна ли данная система справляться с потоком заявок.

Задачи теории массового обслуживания носят оптимизационный характер и в конечном итоге включают экономический аспект по определению такого варианта системы, при котором будет обеспечен минимум суммарных затрат от ожидания обслуживания, потерь времени и ресурсов на обслуживание и

простоев каналов обслуживания.

Система обслуживания считается заданной, если известны:

1) поток требований, его характер;
2) множество обслуживающих приборов;
3) дисциплина обслуживания (совокупность правил, задающих процесс обслуживания).

Каждая СМО состоит из какого-то числа обслуживающих единиц, которые называются каналами обслуживания. В качестве каналов могут фигурировать: линии связи, различные приборы, лица, выполняющие те или иные операции и т.п.

Всякая СМО предназначена для обслуживания какого-то потока заявок, поступающих в какие-то случайные моменты времени. Обслуживание заявок продолжается какое-то случайное время, после чего канал освобождается и готов к приему следующей заявки. Случайный характер потока заявок и времен обслуживания приводит к тому, что в какие-то периоды времени на входе СМО скапливается излишне большое число заявок (они либо становятся в очередь, либо покидают СМО не обслуженными); в другие же периоды СМО будет работать с недогрузкой или вообще простаивать.

Процесс работы СМО представляет собой случайный процесс с дискретными состояниями и непрерывным временем; состояние СМО меняется скачком в моменты появления каких-то событий (или прихода новой заявки, или окончания обслуживания, или момента, когда заявка, которой надоело ждать, покидает очередь).

Перечень характеристик систем массового обслуживания можно представить следующим образом:

- среднее время обслуживания;
- среднее время ожидания в очереди;
- среднее время пребывания в СМО;
- средняя длина очереди;
- среднее число заявок в СМО;
- количество каналов обслуживания;

- интенсивность входного потока заявок;
- интенсивность обслуживания;
- интенсивность нагрузки;
- коэффициент нагрузки;
- относительная пропускная способность;
- абсолютная пропускная способность;
- доля времени простоя СМО;
- доля обслуженных заявок;
- доля потерянных заявок;
- среднее число занятых каналов;
- среднее число свободных каналов;
- коэффициент загрузки каналов;
- среднее время простоя каналов.

Для облегчения процесса моделирования используют классификацию СМО по различным признакам, для которых пригодны определенные группы методов и моделей теории массового обслуживания, упрощающие подбор адекватных математических моделей к решению задач обслуживания в коммерческой деятельности.

СМО с ожиданием — в общем случае многоканальная система в которую поступает поток заявок с интенсивностью λ; интенсивность обслуживания μ (т. е. в среднем непрерывно занятый канал будет выдавать $\frac{\lambda}{\mu}$ обслуженных заявок в единицу (времени). Заявка, поступившая в момент, когда канал занят, становится в очередь и ожидает обслуживания.

Система с ограниченной длиной очереди. Предположим сначала, что количество мест в очереди ограничено числом m, т. е. если заявка пришла в момент, когда в очереди уже стоят m заявок, она покидает систему необслуженной. В дальнейшем, устремив m к бесконечности, мы получим характеристики одноканальной СМО без ограничений длины очереди.

Будем нумеровать состояния СМО по числу заявок, находящихся в системе (как обслуживаемых, так и ожидающих обслуживания):

S_0—канал свободен; S_1—канал занят, очереди нет;

S_2— канал занят, одна заявка стоит в очереди;

S_k—канал занят, k - 1 заявок стоят в очереди;

S_{m+1}— канал занят, k заявок стоят в очереди.

Граф системы показан на рис. 1.2. Все интенсивности потоков событий, переводящих в систему по стрелкам слева направо, равны λ, а справа налево — μ. Действительно, по стрелкам слева направо систему переводит поток заявок (как только придет заявка, система переходит в следующее состояние), справа же налево — поток «освобождений» занятого канала, имеющий интенсивность μ (как только будет обслужена очередная заявка, канал либо освободится, либо уменьшится число заявок в очереди).

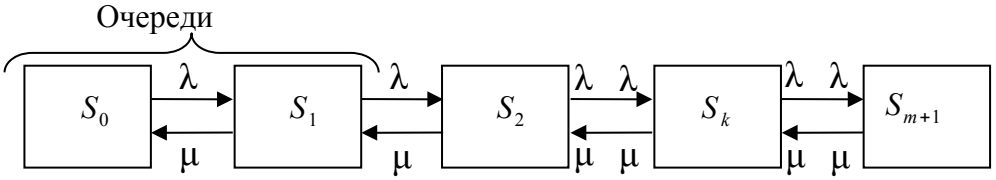

Рис. 1.2. Одноканальная СМО с ожиданием

Изображенная на рис. 1.2 схема представляет собой схему размножения и гибели. Используя общее решение (1.1)—(1. 4), напишем выражения для предельных вероятностей состояний:

$$\begin{cases} p_k = \left(\frac{\lambda}{\mu}\right)^2 p_0; (k = 1, 2, \ldots, m+1); \\ p_0 = \frac{1}{1 + \left(\frac{\lambda}{\mu}\right) + \left(\frac{\lambda}{\mu}\right)^2 + \ldots + \left(\frac{\lambda}{\mu}\right)^{m+1}} \end{cases} \quad (1.1)$$

или с использованием $p_k = \frac{\lambda}{\mu}$:

$$\begin{cases} p_k = p^k p_0; (k = 1, 2, \ldots, m+1); \\ p_0 = \frac{1}{1 + p + p^2 + \ldots + p^{m+1}} \end{cases} = (1 + p + p^2 + \ldots + p^{m+1})^{-1} \quad (1.2)$$

Последняя строка в (1.2) содержит геометрическую прогрессию с первым членом 1 и знаменателем p; откуда получаем:

$$p_0 = \frac{1}{(1 - p^{m+2})/(1 - p)} = \frac{1 - p}{1 - p^{m+2}} \quad (1.3)$$

в связи с чем предельные вероятности принимают вид:

$$\begin{cases} p_0 = \dfrac{1-p}{1-p^{m+2}} \\ p_1 = \rho\, p_0; \\ p_2 = \rho^2\, p_0; \\ \dots \\ p_k = \rho^k\, p_0; \\ p_{m+1} = \rho^{m+1}\, p_0 \end{cases} \qquad (1.4)$$

Выражение (1.4) справедливо только при $\rho < 1$ (при $\rho = 1$ она дает неопределенность вида 0/0). Сумма геометрической прогрессии со знаменателем $\rho = 1$ равна m + 2, и в этом случае

$$p_0 = \frac{1}{m+2}$$

Определим характеристики СМО: вероятность отказа $P_{отк}$, относительную пропускную способность q, абсолютную пропускную способность A, среднюю длину очереди $\bar{r}$, среднее число заявок, связанных с системой $\bar{k}$, среднее время ожидания в очереди $\bar{t}_{ож}$, среднее время пребывания заявки в СМО $\bar{t}_{смо}$.

Вероятность отказа. Очевидно, заявка получает отказ только в случае, когда канал занят и все т мест в очереди тоже:

$$p_{отк} = p_{m+1} = \frac{p^{m+1}(1-p)}{1-p^{m+2}} \qquad (1.5)$$

Относительная пропускная способность:

$$q = 1 - P_{отк} = 1 - \frac{p^{m+1}(1-p)}{1-p^{m+2}} \qquad (1.6)$$

Абсолютная пропускная способность:

$$A = \lambda q$$

Средняя длина очереди. Найдем среднее число $\bar{r}$ заявок, находящихся в очереди, как математическое ожидание дискретной случайной величины R — числа заявок, находящихся в очереди:

$$\bar{r} = M[R]$$

С вероятностью p_2 в очереди стоит одна заявка, с вероятностью p_3 — две заявки, вообще с вероятностью p_k в очереди стоят k - 1 заявок, и т. д., откуда:

$$\bar{r} = M[R] = \sum_{k=2}^{m+1} p^k\, p_0 = ¿\, p^2\, p_0 \sum_{k=2}^{m+1} (k-1)\, p^{k-2} = p^2\, p_0 \sum_{k=1}^{m} k\, p^{k-1} ¿ \qquad (1.7)$$

Поскольку $k\, p^{k-1} = \dfrac{d\, p^k}{dp}$, сумму в (1.7) можно трактовать как производную

по ρот суммы геометрической прогрессии:

$$\sum_{k=1}^{m} k\, p^{k-1} = ¿$$

Подставляя данное выражение в (1.7) и используя p_0из (1.4), окончательно получаем:

$$\bar{r} = \frac{p^2\left(1-\left(m+1-m\rho\right)\rho^m\right)}{\left(1-p\right)\left(1-p^{m+2}\right)} \qquad (1.8)$$

Среднее число заявок, находящихся в системе. Получим далее формулу для среднего числа $\bar{k}$ заявок, связанных с системой (как стоящих в очереди, так и находящихся на обслуживании). Поскольку$\bar{k} = \bar{r} + \bar{w}$, где$\bar{w}$— среднее число заявок, находящихся под обслуживанием, а k известно, то остается определить $\bar{w}$. Поскольку канал один, число обслуживаемых заявок может равняться 0 (с вероятностьюp_0) или 1 (с вероятностью 1 - p_0), откуда:

$$\bar{w} = 0 \cdot p_0 + 1 \cdot \left(1 - p_0\right) = \frac{p - p^{m+2}}{1 - p^{m+2}}$$

и среднее число заявок, связанных с СМО, равно

$$\bar{k} = \bar{r} + \frac{p - p^{m+2}}{1 - p^{m+2}} \qquad (1.9)$$

Среднее время ожидания заявки в очереди. Обозначим его$\bar{t}_{ож}$; если заявка приходит в систему в какой-то момент времени, то с вероятностьюp_0канал обслуживания не будет занят, и ей не придется стоять в очереди (время ожидания равно нулю). С вероятностьюP_1она придет в систему во время обслуживания какой-то заявки, но перед ней не будет очереди, и заявка будет ждать начала своего обслуживания в течение времени$1/\mu$(среднее время обслуживания одной заявки). С вероятностьюP_2в очереди перед рассматриваемой заявкой будет стоять еще одна, и время ожидания в среднем будет равно$2/\mu$, и т. д.

Если же k = m + 1, т. е. когда вновь приходящая заявка застает канал обслуживания занятым и m заявок в очереди (вероятность этогоP_{m+1}), то в этом случае заявка не становится в очередь (и не обслуживается), поэтому время ожидания равно нулю. Среднее время ожидания будет равно:

$$\bar{t}_{ож} = p_1 \frac{1}{\mu} + p_2 \frac{2}{\mu} + \ldots + p_k \frac{k}{\mu} + \ldots + p_m \frac{m}{\mu}$$

если подставить сюда выражения для вероятностей (1.4), получим:

$$\bar{t}_{ож} = p_0\, p\, \frac{1}{\mu} + p_0\, p^2\, \frac{2}{\mu} + \ldots + p_0\, p_k\, \frac{k}{\mu} + \ldots + p_0\, p_m\, \frac{m}{\mu} = \frac{p_0\, p}{\mu}\, \frac{\left(1 - \left(m + 1 - m_p\right) p^m\right)}{\left(1 - p\right)^2} = \frac{p}{\mu}\, \frac{\left(1 - \left(m + 1 - m_p\right) p^m\right)}{\left(1 - p\right)\left(1 - p^{m+2}\right)} \tag{1.10}$$

Здесь использованы соотношения (1.7), (1.10) (производная геометрической прогрессии), а также p_0 из (1.4). Сравнивая это выражение с (1.10), замечаем, что иначе говоря, среднее время ожидания равно среднему числу заявок в очереди, деленному на интенсивность потока заявок.

$$\bar{t}_{ожс} = \frac{1}{p\mu}\, \bar{r} = \frac{\bar{r}}{\lambda} \tag{1.11}$$

Среднее время пребывания заявки в системе. Обозначим $(\bar{t}_{смо})$ матожидание случайной величины — время пребывания заявки в СМО, которое складывается из среднего времени ожидания в очереди $(\bar{t}_{ожс})$ и среднего времени обслуживания $(\bar{t}_{обсл})$. Если загрузка системы составляет 100 %, очевидно, $\bar{t}_{обсл} = 1/\mu$, в противном же случае

$\bar{t}_{обсл} = q/\mu$

Отсюда

$$\bar{t}_{смо} = \bar{t}_{ожс} + \bar{t}_{обсл} = \frac{\bar{r}}{\lambda} + \frac{q}{\mu}$$